A Web-Service Approach for Distributed Access to Methods, Data and Models

(I Don't Care Where My Data and Methods Are)

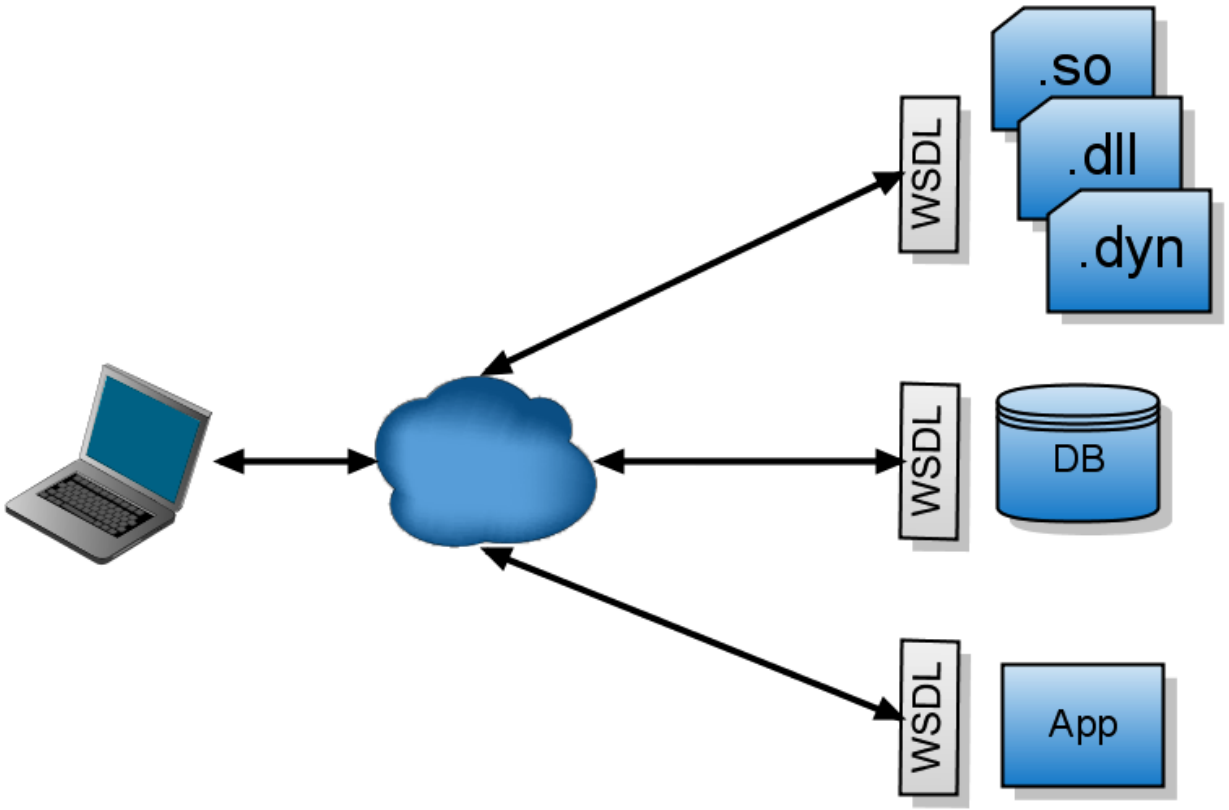
Rajarshi Guha Geoffrey Fox Kevin E. Gilbert
Marlon Pierce David Wild

School of Informatics
Indiana University

235th ACS National Meeting
6th March, 2008

Outline

Local versus Distributed Resources



- Access arbitrary resources (methods, data, applications)
- All resources look like function calls

Combining Data & Functionality

- But what do we mean by "combining" all these resources?
- Different levels of complexity
  - Keep track of new additions to a database via an RSS feed
  - Use Yahoo Pipes to combine and manipulate output easily
  - Write full fledged programs in your language of choice
- Web services allow us to support all these activities in a uniform manner

Web Services - What's Available

Cheminformatics functionality

- Molecular descriptors
- Similarity (2D, 3D)
- Format conversion
- Depictions
- 3D coordinate generation
- Summarized on ChemBioGrid

Dong, X. et al, *J. Chem. Inf. Model.*, **2007**, *47*, 1303–1307

`http://www.chembiogrid.org/projects/proj_core.html`

Web Services for Modeling

- Computational web services can be viewed as wrappers around the actual program
- Since predictive models are a common feature in cheminformatics we'd like to support them as well
- This leads to a number of requirements
  - Ability to develop models
  - Store (deploy) models
  - Use the models via the web service infrastructure
- We provide a computational infrastructure based on R which supports
  - Feature selection
  - Model development
  - Model deployment
  - Arbitrary R code

Guha, R. *J. Chem. Inf. Model.*, **2008**, *48*, 456–464

What's Available?

- Regression (OLS, CNN, RF)
- Classification (LDA)
- Clustering (k-means)
- Feature selection (stepwise and exhaustive)
- Automated model generation
  - Load X, Y data, build linear and non-linear models with optional LOO CV
- Deployed models
  - Random forests for 60 NCI DTP cell lines
  - Cytotoxicity
  - Ames mutagenicity

Wang, H. et al., *J. Chem. Inf. Model.*, **2007**, *47*, 2063–2076

Guha, R. and Schürer, S., *J. Comp. Aid. Molec. Des*, **2008**, ASAP

Web Services - What's Available

Data sources

- We maintain a number of databases
  - PubChem mirror (mainly for local research)
  - Pub3D
- Directly accessible via queries in SQL
- We also encapsulate specific queries as web service calls

http://www.chembiogrid.org/projects/proj_db.html

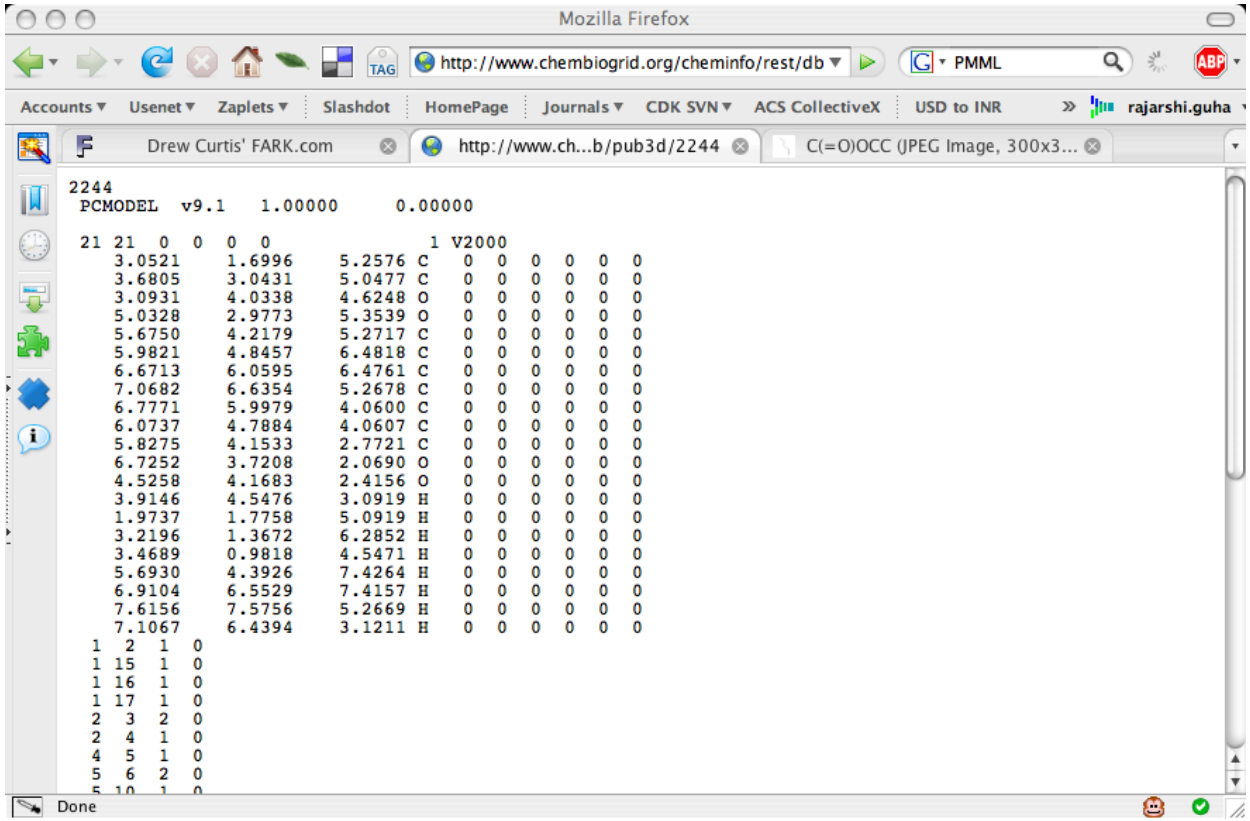
REST Frontends to SOAP Services

- REST is a network architecture that avoids complex message formats
  - No extra libraries required
  - Access services using URL's
  - Much simpler interface compared to SOAP
- We've been putting REST interfaces onto a number of our SOAP services
- Current REST services include
  - Database (3D structure, PubChem mirror)
  - Molecular descriptors
  - Depictions

REST 3D Structure Service

- To get a 3D structure for a PubChem CID

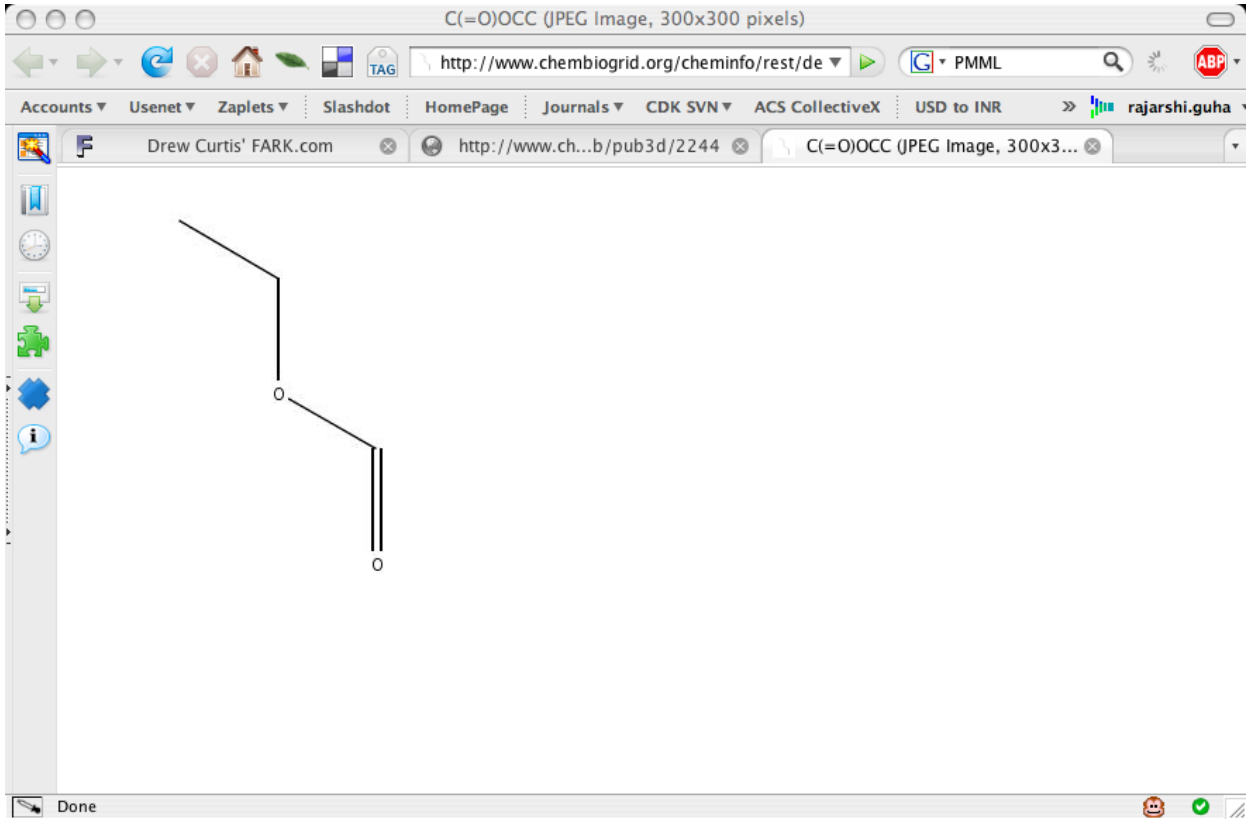
```
http://www.chembiogrid.org/cheminfo/rest/db/pub3d/2244
```



REST Depiction Service

- To get a 2D depiction for arbitrary SMILES

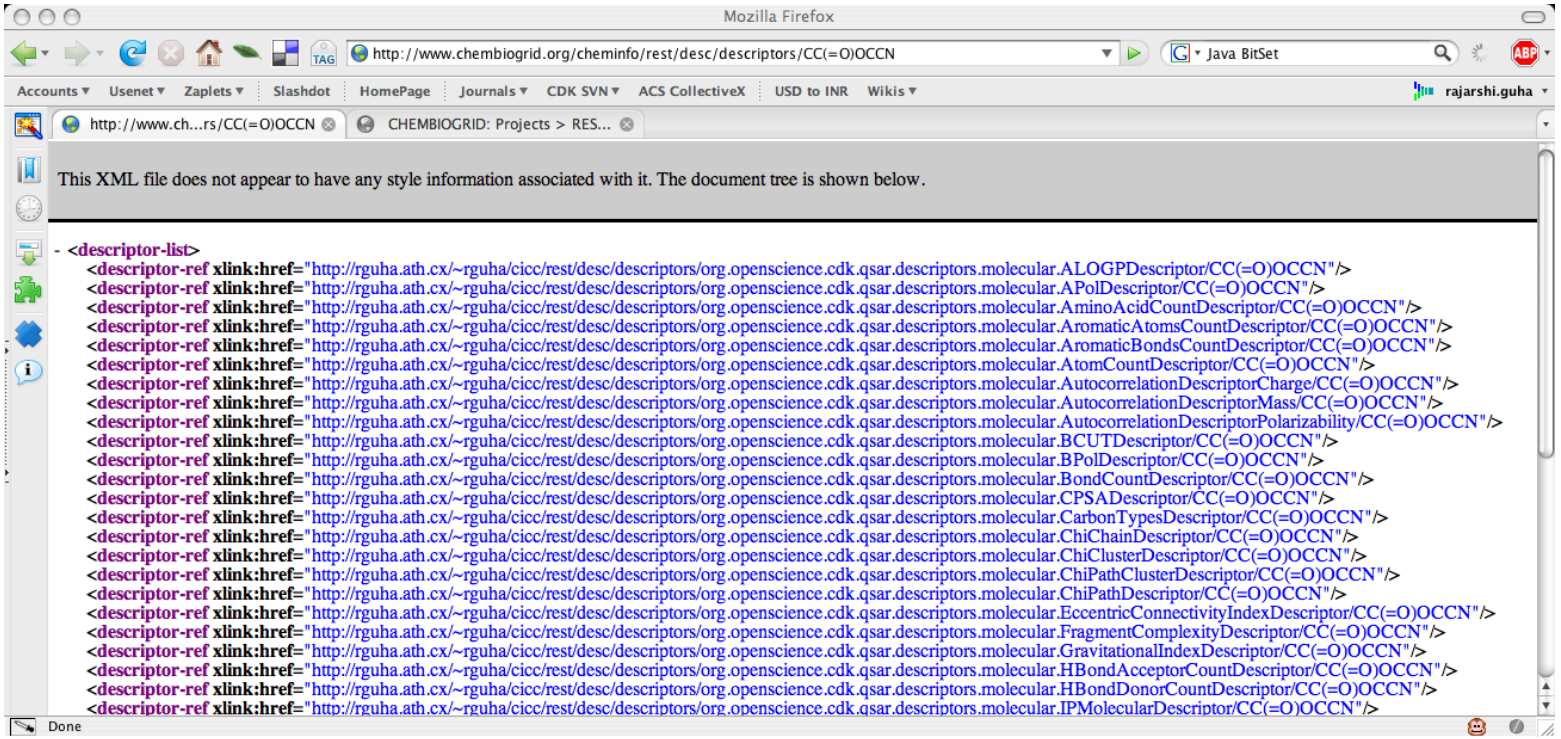
```
http://www.chembiogrid.org/cheminfo/rest/depict/C(=O)OCC
```



REST Descriptor Service

- To get descriptors for an arbitrary SMILES string

```
http://www.chembiogrid.org/cheminfo/rest/desc/
descriptors/CC(=O)OCCN
```



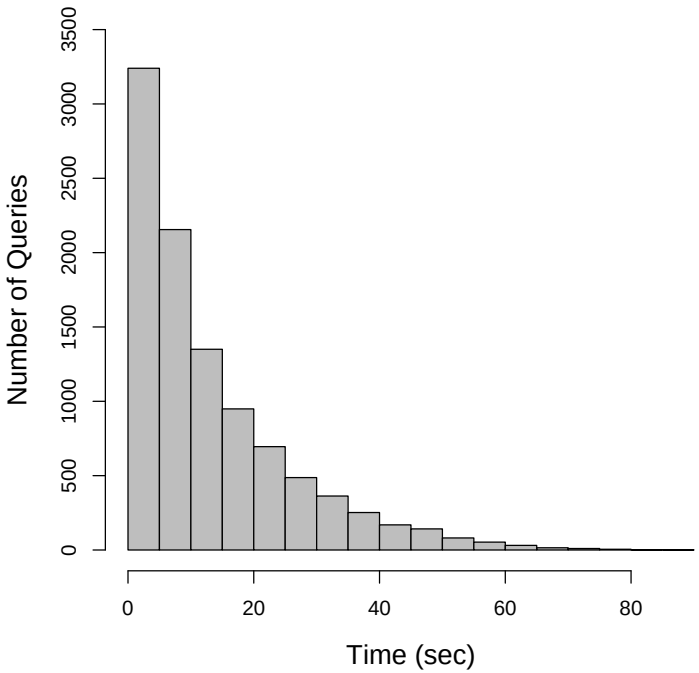
Pub3D

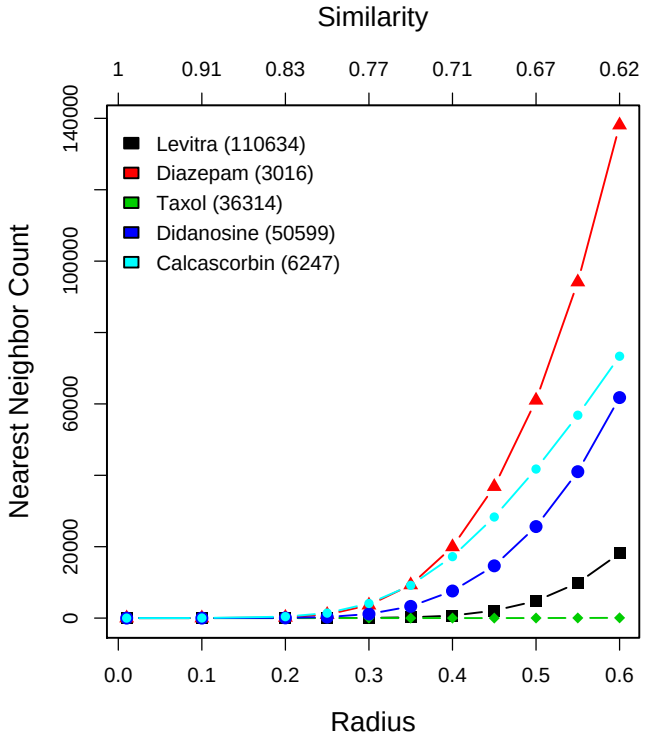
- A 3D structure database derived from PubChem
- Current version contains a single 3D structure for 17M compounds
  - Structures obtained using MMFF94
  - Not the lowest energy conformer
- Structures can be retrieved in SD format by CID using a web page or web service interfaces
- We also store distance moment shape descriptors allowing us to perform shape similarity searches
- http://www.chembiogrid.org/cheminfo/p3d/

Ballester, P.J. and Graham Richards, W., *J. Comp. Chem.*, **2007**, *28*, 1711–1723

Pub3D Performance

- Shape searches can be as fast as 5s, for reasonably large result sets
- Fast enough for us to explore the "density of space" around a given query compound

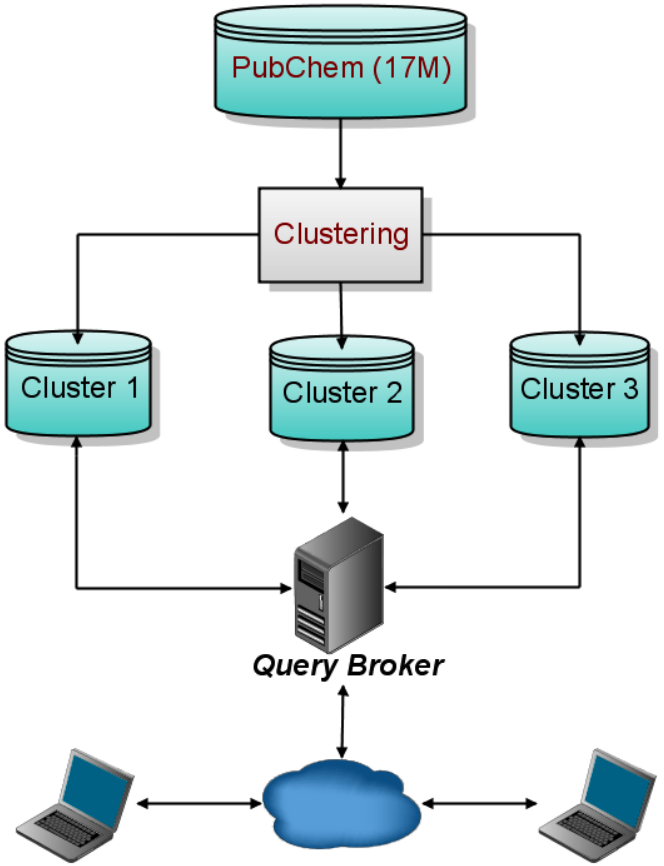




Guha, R. et al., *J. Chem. Inf. Model.*, **2006**, *46*, 1713–1722

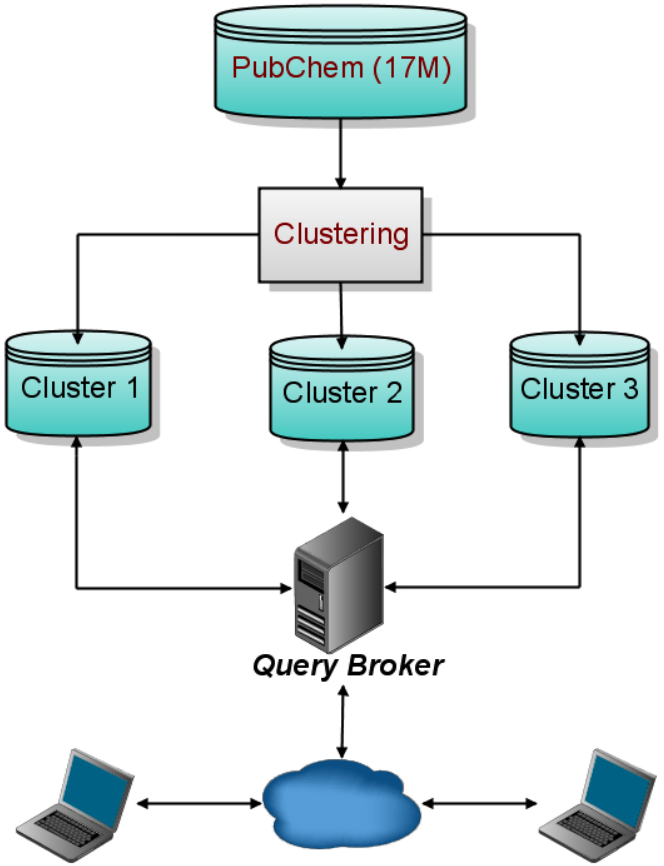
Pub3D & Clustering

- Indexing gives good performance
- As index size increases, performance degrades
- Could add more RAM
- Clustering the database allows us to scale to significantly larger collections



Pub3D & Clustering

- Indexing gives good performance
- As index size increases, performance degrades
- Could add more RAM
- Clustering the database allows us to scale to significantly larger collections



Pub3D and Conformers

- Single, arbitrary conformers aren't too useful
- We have currently generated conformers for a subset of PubChem (4 to 10 heavy atoms)
  - 3 Kcal energy window
  - 243,892 compounds
  - $\approx$ 2M conformers
- Clustering will be vital when conformers are considered
  - Allows us to handle arbitrary numbers of conformers
  - May need to consider some sort of compression
  - Could use cluster information to optimize queries

Model Exchange

- The literature contains many published models
- No way to utilize them unless we manually rebuild them
  - In many cases we will not have access to descriptors
- Difficult to search for models specifically
- We should be able to do the following
  - Search for models
  - Exchange them
  - Execute them
- Is this a "format" issue?
  - To some extent yes

Model Exchange

- The literature contains many published models
- No way to utilize them unless we manually rebuild them
  - In many cases we will not have access to descriptors
- Difficult to search for models specifically
- We should be able to do the following
  - Search for models
  - Exchange them
  - Execute them
- Is this a "format" issue?
  - To some extent yes

PMML for Model Exchange

- **P**redictive **M**odelling **M**arkup **L**anguage
- A standard supported by the Data Modeling Group
- Allows you to serialize predictive models to XML
  - Linear, logistic regression
  - Tree models
  - Neural networks, Naïve Bayes models
  - Association models
  - Ensemble models (random forests, arbitrary ensembles)
- Supported by a number of platforms
  - IBM, Salford Systems
  - SAS, SPSS
  - R

PMML for Model Exchange

- Since it's XML, it's extensible
- PMML allows us to specify the model itself
- But we need to add extra information to make a model truly searchable
  - Provenance (who made it, when was it made, . . .)
  - Property (what is being modeled)
  - Requirements (what are the inputs, how to get them)

PMML for Model Exchange

Provenance

- Easily solved using Dublin Core

Property

- Could be addressed using keywords
- Could reuse pre-existing ontologies

Requirements

- This is tricky
- Need to identify what descriptors were used
- What software, version
- How to evaluate the descriptors (if possible)

PMML for Model Exchange

Provenance

- Easily solved using Dublin Core

Property

- Could be addressed using keywords
- Could reuse pre-existing ontologies

Requirements

- This is tricky
- Need to identify what descriptors were used
- What software, version
- How to evaluate the descriptors (if possible)

PMML for Model Exchange

Provenance

- Easily solved using Dublin Core

Property

- Could be addressed using keywords
- Could reuse pre-existing ontologies

Requirements

- This is tricky
- Need to identify what descriptors were used
- What software, version
- How to evaluate the descriptors (if possible)

Summary

- Pub3D is a shape searchable version of PubChem
- Conformers will have to be considered for it to be useful
- Are searches meaningful? Benchmarks required

- Web services provide one approach to model deployment
- We should be able to search for models explicitly
- PMML is one extensible solution the addresses model exchange
- Automated model execution is more challenging

Summary

- Pub3D is a shape searchable version of PubChem
- Conformers will have to be considered for it to be useful
- Are searches meaningful? Benchmarks required

- Web services provide one approach to model deployment
- We should be able to search for models explicitly
- PMML is one extensible solution the addresses model exchange
- Automated model execution is more challenging

Acknowledgments

- Kangseok Kim
- NIH